

# Extracting Synonymous Gene and Protein Terms from Biological Literature

Hong Yu, Eugene Agichtein

Columbia University

2003

presented by

Hamid Reza Chitsaz

# Names in Biology

## Entities

- Living Creatures: Viruses, Microbes, Species
- Substances: Drugs, Proteins, Lipids, Lipoproteins, Genes, etc.

## Characteristics

- Status, e.g. Clinical Symptoms, and Diseases
- Structure, e.g. Lipoproteins
- Function, e.g. Hormones, Genes, etc.

**Identity, Similarity, and Synonymity?**

# Motivation

## Facts

1. Need of a Synonyms Thesaurus
2. Existence of Large Volume of Growing Biological Literature
3. Cost of Educated Professional Labor

# Problem Assumptions

1. Confinement to Genes and Proteins
2. Synonymity?
3. Co-occurrence in a sentence!
4. Occurrence in the first 4Kb!

# Solutions

## Steps

- Tag Genes and Proteins
- Segment Sentences
- Find Synonymous Pairs, Rank, and List

# Synonymous Pairs: Method (I)

## Unsupervised

### Steps

- For each term  $t_1$  search heuristically for all terms  $t_2$  which are contextually similar to  $t_1$
- Rank and pick top  $k = 5$

# Synonymous Pairs: Method (II)

## Partially-supervised

### Steps

- Start with manual positive and negative pairs
- Extract patterns such as  $\langle GENE \rangle$  *also known as*  $\langle GENE \rangle$
- Find new pairs, Rank, Pick or Dump
- If !happy go to 2

# Synonymous Pairs: Method (III)

## Supervised

### Steps

- Start with manual positive and negative pairs
- Learn patterns as *Positive* or *Negative*
- Find new pairs, Rank, Pick or Dump
- If !happy go to 2

# Synonymous Pairs: Method (IV)

## Hand-crafted(Very-supervised!)

### Steps

- Start with manual positive and negative pairs
- Manually extract patterns
- Find new pairs, Rank, Pick or Dump
- If !happy go to 2

# Synonymous Pairs: Method (V) Combined

## Steps

- Extract synonymous pairs by all previous four methods
- Take concensus

# Evaluation

## Summary

- Unsupervised: Bad Recall, Bad Precision
- Partially-supervised: Good Recall, Bad Precision
- Supervised: Good Recall, Bad Precision
- Hand-crafted: Bad Recall, Good Precision
- Combined: Good Recall, Good Precision  
**i.e. everyone is happy**

# Discussion

- Interesting and useful problem
- Sloppy problem formulation: in particular no synonymy definition
- Hard to reproduce evaluations
- Future directions:
  - Diseases, drugs, symptoms, species, etc.
  - Combine learned patterns with hand-crafted patterns
- Moral: machine learning is helpful